Binomial random variables

---

56 marks

Suppose $X \sim Binomial(n, p)$, then

$$Pr(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \qquad\qquad x = 0, 1, \ldots, n.$$

from which it follows that $E(X) = np$ and $Var(X) = np(1-p)$.

Once x is observed, the unknown proportion can be estimated (e.g. via maximum likelihood) as the numerical value of $\hat{p} = x/n$; the corresponding random estimat**or** would be $\tilde{p} = X/n$.

In this question, you are going to develop your intuition about simple proportions as random variables through a bit of mathematics, some introductory R programming, and a little simulation and visualization.

a. The **odds** (in favour of the event) are defined by the ratio $p/(1-p)$ (e.g. even odds are 1, or 1:1, when $p = 1/2$; odds are 9:1 in favour when $p = 9/10$; odds less than one are often inverted to be described as the odds *against* the event).

   Suppose we are interested in comparing the binomial probabilities $Pr(X = a)$ to $Pr(X = b)$. For example, the ratio, $Pr(X = a)/Pr(X = b)$, tells us how many times more (or less) likely it is to observe $x = a$ than $x = b$.

   i. *(1 mark)* Express the ratio $Pr(X = a)/Pr(X = b)$ as a function of the odds.

   $$\frac{Pr(X=a)}{Pr(X=b)} = \frac{\binom{n}{a} p^a (1-p)^{n-a}}{\binom{n}{b} p^b (1-p)^{n-b}} = \frac{\binom{n}{a}}{\binom{n}{n}} \frac{p^{a-b}}{(1-p)^{-n+a+n-b}} = \frac{\binom{n}{a}}{\binom{n}{b}} \frac{p^{a-b}}{(1-p)^{a-b}} = \frac{\binom{n}{a}}{\binom{n}{b}} odd^{a-b}$$

   ii. *(4 marks)* Here, in two different ways, you will write a function of the odds which calculates $Pr(X = a)/Pr(X = b)$ for a given n.

   For the first way, write `prob_ratio1()` to do the calculation using the function `choose()`:

```
prob_ratio1 <- function (n, a, b, odds = 1) {
   choose_n_a <- choose(n,a)
   choose_n_b <- choose(n,b)
   ans <- choose_n_a/choose_n_b*odds^(a-b)
   return(ans)
}
# print(prob_ratio1(5,3,4,5))
```

```
For the second way, write `prob_ratio1()` to do the calculation using the function `dbinom()`:
```

```
prob_ratio2 <- function (n, a, b, odds = 1) {
  p <-  odds/(1+odds)
  p_a <- dbinom(a,size=n,prob=p)
  p_b <- dbinom(b,size=n,prob=p)
  return(p_a/p_b)
}
# print(prob_ratio2(5,3,4,5))
```

```
Both are calculating the same values.
```

```
iii. *(2 marks)* Report the following values for `prob_ratio1()` and `prob_ratio2()`
```

```r
# using choose()
prob_ratio1(50, a = 5, b = 45)
```

```
## [1] 1
```

```r
prob_ratio1(50, a = 5, b = 45, odds = 9)
```

```
## [1] 6.765496e-39
```

```r
# and using dbinom()
prob_ratio2(50, a = 5, b = 45)
```

```
## [1] 1
```

```r
prob_ratio2(50, a = 5, b = 45, odds = 9)
```

```
## [1] 6.765496e-39
```

b. Extreme proportions like $\widehat{p} \approx 0$ or $\widehat{p} \approx 1$ often generate a great deal of interest in an analysis.

   For example, p might be the proportion of people in some population who perhaps die from some exposure to some toxin, or are cured of a disease by some treatment, or maybe just say they would vote for a particular party or candidate. In any of these cases it can be surprising (even alarming) to see either $\widehat{p} \approx 0$ or $\widehat{p} \approx 1$, so much so that some explanation seems in order.

   Suppose we have observed x from $Binomial(n, p)$ and y from $Binomial(m, p)$ – that is the same probability of occurrence but different sample sizes. Denote the observed proportions as $\widehat{p}_x = x/n$ and $\widehat{p}_y = y/m$, respectively.

   i. *(2 marks)* Give the mathematical expression for the ratio

   $$\frac{Pr(\widetilde{p}_x = 0)}{Pr(\widetilde{p}_y = 0)}$$

   and for the ratio

   $$\frac{Pr(\widetilde{p}_x = 1)}{Pr(\widetilde{p}_y = 1)}$$

   .

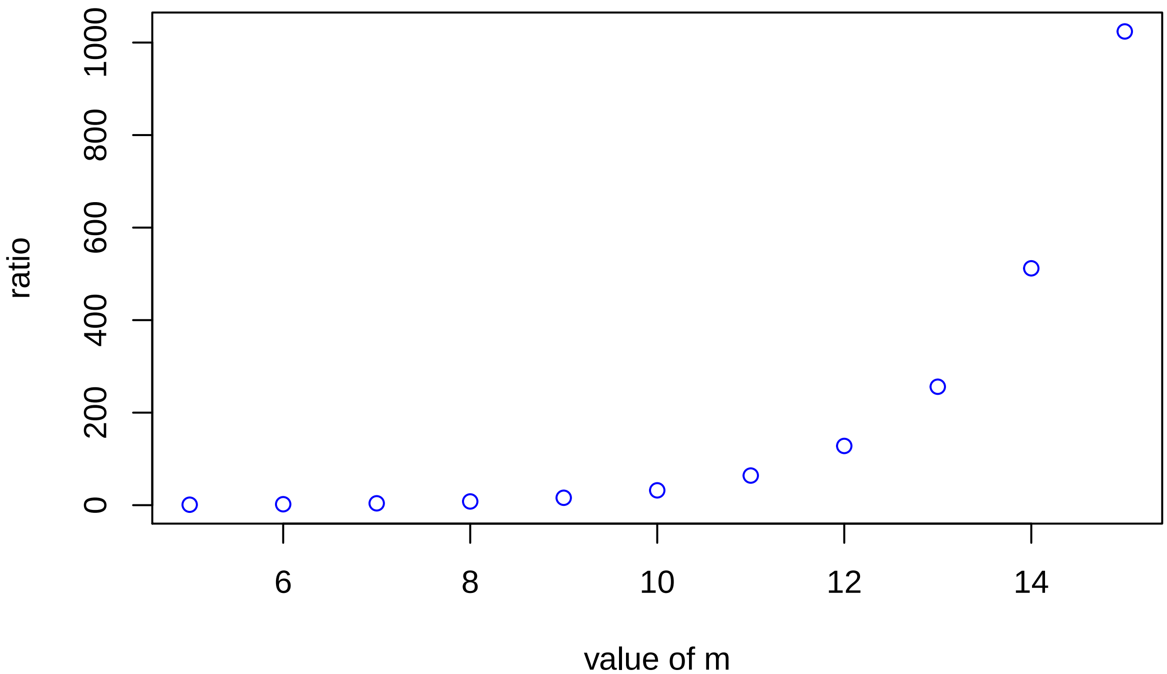
   $$\frac{Pr(\widetilde{p}_x = 0)}{Pr(\widetilde{p}_y = 0)} = \frac{\binom{n}{0}p^0(1-p)^n}{\binom{m}{0}p^0(1-p)^m} = \frac{\binom{n}{0}}{\binom{m}{0}}(1-p)^{n-m} = (1-p)^{n-m}$$

   $$\frac{Pr(\widetilde{p}_x = 1)}{Pr(\widetilde{p}_y = 1)} = \frac{\binom{n}{n}p^n(1-p)^0}{\binom{m}{m}p^m(1-p)^0} = \frac{\binom{n}{n}}{\binom{m}{m}}p^{n-m} = p^{n-m}$$

   ii. *(2 marks)* Describe what happens to each of these ratios for $n < m$ and m increases. As m increases, the first ratio will increase. The base 1-p is between 0 and 1 and the exponent decreases from 0. The power will increase if m increases. As m increases, the second ratio will increase. The base p is between 0 and 1 and the exponent decreases from 0. The power will increase if m increases.

   iii. *(3 marks)* Suppose $p = 1/2$, $n = 5$ and $m \in \{5, 6, 7, \ldots, 14, 15\}$. Using the `plot()` function (with appropriate title and axis labels), plot the curve of the pair (n, m) for all values.

```r
ms <- seq(from=5, to=15)
get_ratio <- function(p,n,ms){
    ans = c()
    for (i in 1:length(ms)){
        ans <- c(ans, p^(n-ms[i]))
    }
    return(ans)
}
ratios <- get_ratio(1/2,5,ms)
plot(ms,ratios,xlab="value of m",ylab="ratio",col="blue")
```




```
iv. *(2 marks)* In plain English, express when $\widehat{p}$ is most likely to be 1 and when you should
```

c. For the binomial definition, as given above:

   i. *(2 marks)* Mathematically derive $E(\tilde{p})$ and the **standard deviation** $SD(\tilde{p})$.

   ii. *(2 marks)* Write a function that calculates the standard deviation of $\tilde{p}$ for any pair of values of n and p as follows:

```r
sd_p_wig <- function (n, p) {
    # Your code here
}
```

d. Chebyshev's inequality relates the nearness of a random variable Y to its expectation μ as a function of its standard deviation σ (provided both exist and are finite) as follows:

$$Pr\left(|Y-\mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

for any constant $k > 1$.

   i. *(2 marks)* Write this inequality when $Y = \tilde{p}$, the binomial proportion estimator.

ii. *(4 marks)* Suppose that $k = 5$. As a function of p, mathematically express the sample size n needed to ensure by Chebyshev's inequality that for our estimator $\tilde{p}$ we have

$$Pr\left(|\tilde{p} - p| \geq \frac{1}{50}\right) \leq \frac{1}{25}.$$

Using the function `plot()`, with appropriate title and x and y axis labels, plot the curve of n as a function of p for `p = seq(0,1, 0.01)`.

iii. *(4 marks)* Suppose that $k = 5$, and $n = 2500$. As a function of p, mathematically express the bound B given by Chebyshev's inequality, so that for our estimator $\tilde{p}$ we have

$$Pr\left(|\tilde{p} - p| \geq B\right) \leq \frac{1}{25}.$$

Using the function `plot()`, with appropriate title and x and y axis labels, plot the curve of B as a function of p for `p = seq(0,1, 0.01)`.

iv *(2 marks)* In simple English, summarize what the largest B says about how well p is likely to be estimated when $n = 2500$ according to Chebyshev's inequality.

e. The functions `plot()`, `lines()`, and `abline()` can be used to plot some data, to add curves, and to add a straight line to a plot. To learn more, type `help("plot.default")`, `help("lines")`, and `help("abline")`.

Use these plotting functions, and appropriate arguments (including meaningful titles, and axis labels), to produce the following plots of the function `sd_p_wig()`:

i. *(3 marks)* As a function of `p = seq(0, 1, 0.01)` for a fixed `n = 10`. Add a vertical **dashed** line in **red** at the value of p which maximizes the standard deviation.

Hand in your code and plot.

ii. *(4 marks)* As a function of `n` $\in \{5, 10, 15, ..., 50\}$ for the different values of `p` $\in \{0.1, 0.3, 0.5, 0.8\}$.

- Use a different colour and line type for each curve (i.e. value of `p`).
- Use `lwd = 2` for all curves.
- Use `legend()` to add a legend to the "'topright" corner of the plot, identify each curve by its colour and line type.
- Hand in your code and plot.

iii. *(2 marks)* Comment on your findings about the dependency of the standard deviation of the binomial proportion estimator $\tilde{p}$ as a function of n and p.

f. The R function `rbinom()` can be used to generate pseudo-random values `x` from a binomial distribution. In this question, you will examine the samples of the sample **proportions**, x/n, drawn from binomials with the same value of p but different values of n.

i. *(3 marks)* Using `rbinom()`, for every $n \in \{5, 10, 15, ..., 45, 50\}$ generate 100 proportions, each one based on an independent pseudo-random value from $Binomial(n, 0.5)$.

You might find the `rep()` function useful in constructing `n`.

Show your code.

ii. *(3 marks)* Plot the pairs $(n, \hat{p})$ as points (use plot arguments `ylim = c(0,1), pch = 19, col = adjustcolor("steelblue", 0.2), cex = 0.5`).

Note that there will be 100 proportions for every n.

Add a red dashed horizontal line at $p = 0.5$.

iii. *(2 marks)* Repeat the production of the above plot, complete with horizontal line, but instead of using `n` as the x variable in the plot, use `jitter(n, 2)`.

   Show your code and resulting plot.

   Comment on the effect of `jitter()`

iv. *(1 mark)* Based on the either of the above plots, what do you conclude about the distribution of binomial proportions as n increases?

g. *(6 marks)* With supporting reference to any/all suitable discoveries you have made in the above questions, comment on each of the following:

   i. Which values for the true binomial probability p are hardest/easiest to estimate from a sample. Why?

   ii. Law of large numbers? What can you say about the effect of increasing sample size n on the quality and/or interpretation of your estimate $\hat{p}$ of p?

   iii. Law of small numbers? What can you say about the effect of decreasing sample size n on the quality and/or interpretation of your estimate $\hat{p}$ of p?